

An Evidence-Derived Framework for Quantifying Translational Risk in Rare-Disease Drug Development

Hanyu Su

hsu9.su@gmail.com

Lakeside High School

Research Article

Keywords: rare diseases, orphan medicines, drug development, translational research, Translation Risk Score, Phase III advancement, evidence synthesis

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11152609/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

An Evidence-Derived Framework for Quantifying Translational Risk in Rare-Disease Drug Development

Hanyu Su^{1*}

^{1*}Independent Researcher, Augusta, GA, USA.

Corresponding author(s). E-mail(s): hsu9.su@gmail.com;

Abstract

Rare diseases differ substantially in the biological evidence, clinical infrastructure, regulatory precedent, economic activity, and research networks available to support therapeutic development. We developed POROS, an auditable cross-disease framework that integrates prespecified machine-observable variables with provenance-linked documentary evidence across five translational domains to generate a Translation Risk Score (TRS; 0–100, with higher values indicating greater modeled risk). Historical feature states were frozen at 31 December 2015 for 100 rare diseases, and Phase III advancement, defined as initiation of at least one Phase III trial, was assessed through 31 December 2025. The documentary evidence extractor achieved 0.873 accuracy and $F1 = 0.838$ across 332 evaluable decisions reviewed by Reviewer 1. In an independent 85-item second-annotator sample, machine–Reviewer 2 Cohen’s κ was 0.616. Cross-validated AUC was 0.822 for TRS, 0.771 for Evidence Coverage, 0.837 for TRS plus Evidence Coverage, and 0.838 for the five-domain model. Phase III advancement declined from 100% in the lowest-TRS quintile to 30% in the highest. TRS and Evidence Coverage were highly correlated ($r = -0.941$). POROS provides a reproducible framework for comparing translational readiness across rare diseases and identifying evidence gaps associated with later-stage therapeutic development. Interpretation remains limited by documentation dependence, modest sample size, and non-causal counterfactual analyses, and external validation is needed before clinical or regulatory decision use.

Keywords: rare diseases, orphan medicines, drug development, translational research, Translation Risk Score, Phase III advancement, evidence synthesis

1 Introduction

Rare-disease drug development often progresses unevenly across the steps required to move from biological insight to a therapy that can be tested in patients. A disease may have a strong mechanistic rationale while still lacking the patient cohorts, natural-history data, outcome measures, regulatory precedent, or organizational capacity needed for late-stage development. Small and geographically dispersed populations add further difficulty to recruitment and trial execution [1–3]. These constraints interact, so progress in one area may not compensate for a bottleneck elsewhere. Orphan-drug development has expanded substantially over recent decades, but development activity and approvals remain uneven across rare diseases [4–6].

These field-level differences are consequential for rare diseases because the maturity of the surrounding evidence ecosystem can determine whether mechanistic discoveries can be translated into evaluable interventions. A cross-disease framework that makes these differences explicit may help distinguish diseases with strong biological knowledge but limited clinical or organizational readiness from those with more mature development pathways [2, 3].

Trial readiness has therefore become a recurring focus for regulators and research agencies. Natural-history studies, patient registries, biomarkers, outcome assessments, and related infrastructure can determine whether a promising therapeutic concept can be evaluated rigorously [3, 7, 8]. Existing readiness frameworks describe important aspects of research and development maturity [2, 3]. POROS was developed to complement this work by representing distinct sources of translational risk with reproducible, source-linked variables while keeping evidence completeness separate from modeled risk. Recent work has also proposed a Disease Readiness Level (DRL) framework that grades disease understanding and product-development progress across four rare-disease case studies [9]. POROS extends this line of work by quantifying multidomain evidence across a 100-disease cohort, freezing predictors before longitudinal follow-up, and testing their association with subsequent Phase III initiation.

POROS addresses three questions: (1) which observable characteristics of a rare-disease field are associated with translational risk; (2) which translational domain forms the dominant modeled bottleneck for a disease; and (3) among prespecified modifiable features, which counterfactual changes produce the largest modeled improvement under the frozen model. Disease-level scores are generated from the prespecified evidence architecture rather than assigned by investigator impression.

This study describes the POROS architecture, evaluates the documentary evidence extractor, applies a temporally separated historical validation design to test whether disease-level evidence states are associated with subsequent Phase III advancement, and defines Counterfactual Translation Risk (CTR) analysis. Visceral myopathy is retained as a disease-specific case study because its defined smooth-muscle biology coexists with substantial clinical-development and infrastructure constraints.

2 Materials and methods

2.1 Study design and analytic architecture

POROS was designed as a cross-disease observational and modeling framework. The analytic workflow consists of: disease identity resolution; retrieval of structured and documentary evidence; feature-level classification with provenance; construction of historical and contemporary disease snapshots; cohort-relative normalization; calculation of five domain scores and the composite TRS; evidence-completeness analysis; historical outcome ascertainment; predictive model evaluation; and counterfactual scenario analysis. Feature definitions, direction of risk, normalization rules, modifiability status, and historical outcome logic are frozen for the manuscript analysis rather than adjusted after inspection of final outcomes. The study design and analytic workflow are summarized in Figure 1.



Fig. 1 POROS study design and analytic workflow. Historical evidence for a frozen cohort of 100 rare diseases was captured using 17 prespecified, provenance-linked evidence features. Features were aggregated into five translational domains: Biological, Clinical, Regulatory, Economic, and Infrastructure, and combined into the Translation Risk Score (TRS; 0–100), with higher values indicating greater modeled translational risk. The historical outcome was Phase III advancement during follow-up, defined as initiation of at least one Phase III trial. Validation and interpretation analyses included cross-validated discrimination, TRS quintile and outcome-group comparisons, extractor validation, independent second-annotator reliability, and counterfactual feature analysis.

2.2 Evidence-variable classes and objective coding

Every manuscript variable belongs to one of three classes. Type A variables are machine-observable quantities obtained from structured sources, including trial counts, dates, phases, sponsor activity, publication or funding measures where implemented, and other reproducibly retrievable quantities. Type B variables are documentary features classified under prespecified definitions, such as evidence of a natural-history study, patient registry, validated outcome measure, disease model, preclinical rescue, or patient organization. Type C variables are mathematically derived quantities, including normalized feature risks, domain scores, TRS, evidence-completeness measures, fitted probabilities, and counterfactual changes.

POROS prohibits direct investigator assignment of disease-level risk scores. Type B classifications represent documented evidence under the prespecified retrieval and classification protocol. A status of NOT CONFIRMED therefore means that qualifying evidence was not confirmed by that protocol; it does not establish that the resource, study, or property is absent in the real world. Unknown or unascertained information is not silently converted to a negative finding.

2.3 Disease resolution, evidence retrieval, and provenance

The platform accepts free-text disease terms and resolves them to a canonical disease identity and associated synonyms before evidence retrieval. Structured evidence is acquired from public biomedical sources used by the implemented pipeline, including ClinicalTrials.gov, PubMed/Europe PMC, NIH RePORTER, and regulatory sources where reproducible retrieval is technically supported. ClinicalTrials.gov provides the primary structured trial-registry source used in the historical analyses [10]. Documentary retrieval is disease-specific and evaluated under the frozen Type B feature definitions. Each raw observation and documentary decision is stored with disease identity, snapshot date, feature identifier, source or document identifier, retrieval time, query context, and supporting evidence where available. This provenance layer permits a score or classification to be traced back to the evidence used to generate it. Prototype modeled baseline values used in earlier interface development are excluded from manuscript analyses.

2.4 Type B documentary evidence extraction

The manuscript implementation uses the frozen Type B ruleset v3.1. Documentary features are evaluated using feature-specific semantic conditions designed to require qualifying evidence rather than isolated keyword occurrence. The classifier is intended to identify documented evidence that satisfies each feature definition within disease-relevant text. Examples include requiring prospective use for biomarker-in-trial evidence, validation evidence tied to the relevant outcome measure for validated-outcome-measure status, and a disease model, intervention, and improvement or rescue relationship for preclinical-rescue status. The extractor preserves the distinction between CONFIRMED PRESENT and NOT CONFIRMED. Its purpose is high-throughput, reproducible evidence classification across the cohort, with human review used to estimate classification performance rather than to manually overwrite the complete corpus.

2.5 Biological tractability domain

Biological tractability represents the degree to which available evidence supports conversion of disease biology into a therapeutic strategy. Documentary features include causal molecular basis, actionable therapeutic target, human-relevant disease model, in vivo model, and preclinical rescue, together with other frozen biological features included in the manuscript feature dictionary. Each documentary classification requires qualifying source evidence under the corresponding frozen rule.

2.6 Clinical development domain

Clinical development risk represents the capacity to characterize disease progression, identify and recruit patients, measure treatment effects, and execute informative interventional studies. Features include natural-history evidence, patient registries, validated outcome measures, prospective biomarker use in trials, and structured

ClinicalTrials.gov-derived measures such as trial activity, phase, enrollment, and recruiting breadth where available in the frozen feature set [1, 3, 7].

2.7 Regulatory pathway domain

Regulatory pathway risk represents the amount of relevant regulatory precedent and evidentiary clarity available to a development program. Frozen features include reproducibly ascertainable evidence of orphan designation, expedited designation, surrogate-endpoint precedent, prior approval or related regulatory precedent where supported by the implemented sources. Best-effort regulatory signals are identified as such and are not represented as authoritative approval-history ascertainment when the underlying public interface does not support that claim [8].

2.8 Economic sustainability domain

Economic sustainability is operationalized using observable development-ecosystem activity rather than investigator estimates of profitability or commercial attractiveness. Implemented measures can include commercial sponsor activity, industry-sponsored interventional trials, distinct development programs, federal research activity, patient-population evidence where defensibly sourced, and sponsor concentration. The framework does not assign subjective return-on-investment scores.

2.9 Translation infrastructure domain

Translation infrastructure captures shared organizational and research capacity supporting development. Frozen documentary features include patient organizations, registries, biobanks or biospecimen resources, natural-history consortia, multicenter networks, and consensus guidance or related coordinated infrastructure where defined in the feature dictionary. These variables are coded only when the retrieved evidence satisfies the prespecified definition.

2.10 Historical snapshots and leakage control

The primary historical design reconstructs disease evidence as of 31 December 2015. Evidence used to construct historical predictors is restricted by the applicable snapshot logic so that later information does not define earlier feature states. Historical ClinicalTrials.gov features use dates such as study start, posting, or completion as appropriate rather than present-day trial status [10]. Follow-up is evaluated separately through 31 December 2025. The primary automated historical outcome is subsequent Phase III initiation because it is more reproducibly timestamped from trial records than a best-effort openFDA label signal. Regulatory label signals may be retained as supplementary information but are not treated as authoritative approval-date history. Outcome ascertainment (trial-registry lookup against ClinicalTrials.gov) and TRS computation (documentary and structured feature-evidence lookup) are mechanically independent pipelines operating on disjoint data paths; no step in outcome derivation consults a computed TRS, domain score, or feature risk value, and no step in score computation consults the outcome table. This provides de facto blinding of outcome

ascertainment to TRS by construction, though it was not additionally enforced by a separate human-blinding protocol since both processes are fully automated.

2.11 Cohort-relative normalization and Translation Risk Score

Continuous Type A variables are transformed when needed for skew and normalized against the frozen manuscript cohort using empirical cohort-relative distributions rather than investigator-selected cut points. Feature direction is prespecified so that larger normalized values consistently represent greater modeled translational risk. Binary documentary features are mapped according to their frozen direction of risk. For the transparent descriptive model, normalized features are aggregated within their assigned domains and domain scores are expressed on a 0 to 100 risk scale. The composite TRS is the equal-weight mean of the five domain scores, giving each domain 20% weight in the primary descriptive score. Equal weighting is a transparency choice and is not claimed to be statistically optimal. Fitted predictive models are evaluated separately.

2.12 Evidence completeness and missing evidence

POROS reports evidence completeness separately from translational risk. Ascertainment completeness describes the proportion of required fields for which a usable value or classification was obtained, while documentary evidence coverage describes the extent to which qualifying positive Type B evidence was confirmed. These quantities answer different questions and are not substituted for one another. Missing or unascertained evidence is not automatically coded as absence. Model-fitting procedures use only the prespecified handling strategy for missing values, and sensitivity analyses may restrict analyses to diseases meeting defined completeness thresholds. This separation is important because sparse documentation can otherwise create an artificial appearance of either high or low readiness. In the frozen manuscript cohort, Ascertainment Completeness itself varied minimally (100.0 for 95/100 diseases, 96.4 for the remaining 5; mean 99.82, SD 0.79), so a completeness-threshold sensitivity analysis was not informative here and is not reported as a separate result; this is noted rather than silently omitted.

2.13 Human validation of the type B extractor

Extractor performance was evaluated using a locked, deterministic, feature-stratified human-validation sample drawn from eligible Type B evidence decisions generated by ruleset v3.1. The completed audit contained 340 human-reviewed classifications across 17 documentary features (20 per feature), all at the 31 December 2015 historical snapshot used in the cohort analysis. The primary reviewer assigned CONFIRMED PRESENT, NOT CONFIRMED, or AMBIGUOUS (reserved for cases the reviewer could not confidently resolve from the retrieved evidence) using the frozen feature definitions and the retrieved supporting material. Machine classifications were compared with the primary reviewer labels without using the human audit to retroactively redefine the evaluated predictions. The 8/340 (2.4%) AMBIGUOUS-labeled determinations were excluded from the primary accuracy/precision/recall/F1/kappa

calculation reported below; a sensitivity calculation including them (coded as NOT CONFIRMED) is also reported and does not change the conclusion. Overall extractor performance was summarized using accuracy, precision, recall, F1 score, and Cohen’s kappa [11], each with 95% confidence intervals from 2,000-replicate bootstrap resampling [12]. The primary extractor metrics reflect agreement between the machine extractor and Reviewer 1 and should not be interpreted as inter-rater reliability for the underlying evidence constructs. To estimate human inter-rater reliability, an independent second annotator reviewed a locked 85-item stratified subset (5 items per feature across all 17 features) while blinded to both machine and Reviewer 1 labels. Reviewer–reviewer and machine–Reviewer 2 agreement were summarized using raw agreement and Cohen’s kappa; ambiguous labels were excluded pairwise from kappa calculations.

2.14 Statistical analysis and historical validation

The frozen analysis dataset was used to summarize cohort characteristics, evidence completeness, domain-score distributions, and relationships among domains. The primary historical endpoint was Phase III initiation during the prespecified follow-up period. Associations between TRS or domain-level predictors and the outcome were evaluated using logistic regression where event counts permitted [13].

Predictive discrimination was evaluated using receiver operating characteristic area under the curve (AUC) [14, 15]. The implemented pipeline used cross-validated predicted probabilities, with stratified k-fold cross-validation when class counts permitted and leave-one-out validation for very small classes. AUC confidence intervals were estimated by bootstrap resampling [12]. Probabilistic performance was additionally summarized using the Brier score [15, 16]. Separate calibration slopes and intercepts were not estimated in this internal validation analysis. The equal-weight descriptive TRS was compared with fitted models without redefining the descriptive score after outcome inspection.

Given the conceptual overlap between TRS and documentary Evidence Coverage, their Pearson and Spearman correlations and variance inflation factors (VIF) were computed directly, and pairwise model discrimination differences (TRS vs. Evidence Coverage; TRS+Evidence Coverage vs. TRS alone; five-domain model vs. TRS alone) were assessed via 1,000-replicate bootstrap resampling of the AUC difference, reporting the mean difference and percentile 95% CI. Because an unpenalized multivariable logistic fit of the five domains showed quasi-complete separation (coefficient standard errors on the order of 10^5 to 10^7 for at least one domain), domain-level coefficients are reported from an L2-penalized ridge logistic model [17], with standardized predictors and 1,000-replicate bootstrap 95% CIs, rather than from the unstable unpenalized maximum-likelihood fit. Separation is a recognized source of instability in logistic regression [18].

2.15 Counterfactual Translation Risk

CTR is a model-based scenario analysis over prespecified modifiable features. A counterfactual changes one eligible factual input at a time, or a clearly specified combination of eligible inputs, while holding all other observed features fixed. The altered

feature vector is passed through the same frozen descriptive or fitted model used for the observed state. The evidence database itself is not mutated by a counterfactual scenario. For descriptive analyses, the scenario effect is reported as the change in TRS. For fitted probability models, it is reported as the change in predicted probability of the prespecified translational outcome. Scenarios are ranked by modeled improvement. CTR outputs are associations implied by the frozen model and are not causal estimates of what would occur if an organization implemented the corresponding intervention [19].

2.16 Visceral myopathy case study

Visceral myopathy was prespecified as a detailed case study because defined smooth-muscle disease mechanisms coexist with substantial clinical-development and infrastructure constraints [20, 21]. The case analysis uses the same evidence definitions and scoring architecture as the cohort analysis, identifies the dominant modeled bottleneck, and evaluates eligible counterfactual scenarios using the same provenance-linked evidence inputs. Input from disease experts or patient organizations is treated as scientific consultation unless collected under an independently reviewed human-subjects research protocol.

2.17 Software, versioning, and reproducibility

POROS is implemented as a reproducible software pipeline that persists disease identities, evidence, snapshots, feature decisions, scores, outcomes, counterfactuals, and analysis runs. The manuscript analysis uses the frozen feature dictionary and Type B extractor ruleset v3.1. Source code for the manuscript-matched software release is archived in Zenodo at <https://doi.org/10.5281/zenodo.22168470> and maintained at <https://github.com/thegrapeslayer/poros-platform>. The frozen disease-level dataset, feature dictionary, and derived statistical and counterfactual-analysis outputs are archived separately in Zenodo at <https://doi.org/10.5281/zenodo.22168501>. Original third-party source documents are not redistributed.

3 Results

3.1 Frozen cohort characteristics and evidence completeness

The frozen manuscript cohort comprised 100 rare diseases (cohort ID manuscript_99262879ff); all were successfully resolved and scored with zero retrieval errors. Phase III advancement was observed for 76 diseases (76.0%) and not observed for 24 (24.0%) over the 2016-01-01 to 2025-12-31 follow-up window. Ascertainment Completeness was near-uniform across the cohort (mean 99.82, SD 0.79; only two distinct values were observed, 100.0 for 95/100 diseases and 96.4 for the remaining 5). This variable is reported descriptively only; its near-constant value across the cohort means it carries little discriminative information and was not included as a covariate in any model below. Documentary Evidence Coverage, in contrast, varied substantially across the cohort (mean 32.7, median 32.35, IQR 17.6–47.1, range 0–70.6),

reflecting genuine cross-disease heterogeneity in how much confirmable documentary evidence the retrieval protocol located, independent of Ascertainment Completeness.

3.2 Type B extractor validation

All 340 prespecified human-validation decisions were reviewed (20 per feature across 17 documentary features, entirely at the 2015-12-31 historical snapshot). Excluding 8 (2.4%) determinations the reviewer labeled AMBIGUOUS, the Type B extractor achieved 0.87 accuracy (95% CI 0.84–0.91), 0.78 precision (95% CI 0.72–0.85), 0.90 recall (95% CI 0.85–0.95), $F1 = 0.84$ (95% CI 0.79–0.88), and Cohen’s kappa = 0.74 (95% CI 0.66–0.81) against the single human reviewer’s labels ($n=332$). Including the 8 AMBIGUOUS determinations, recoded as NOT CONFIRMED, changed the pooled metrics only marginally (accuracy 0.8706, precision 0.773, recall 0.9008, $F1 = 0.8321$, kappa 0.7278; $n=340$) — the handling of AMBIGUOUS cases did not materially affect the reported performance. Performance varied by feature, indicating that some documentary constructs were substantially easier to classify reproducibly than others. Feature-level $F1$ ranged from 0.588 for consensus guidance to 1.000 for expedited-designation evidence, in vivo model evidence, and patient-organization evidence. Table 1 reports precision, recall, and $F1$ for all 17 features. Lower-performing features (consensus guidance, $F1=0.588$; multicenter network, $F1=0.625$; surrogate endpoint precedent, $F1=0.727$) are reported explicitly rather than concealed by the pooled metric and should be considered when interpreting disease-level scores — particularly the Infrastructure and Regulatory domains, which draw on several of them — that depend on those constructs.

Reviewer 1’s labels were used as the primary reference standard for extractor validation; the independent second-annotator analysis is reported below. In the independent 85-item dual-annotator subset, all three evaluators (Reviewer 1, Reviewer 2, and the machine extractor) were in complete agreement on 59 of 85 items (69.4%). Both human reviewers agreed while the machine differed in only 4 cases (4.7%). Disagreements between human annotators were distributed between Reviewer 1 matching the machine while Reviewer 2 differed (12/85, 14.1%) and Reviewer 2 matching the machine while Reviewer 1 differed (8/85, 9.4%), with 2 items (2.4%) involving an AMBIGUOUS designation. Evaluated against Reviewer 2’s labels as an independent secondary benchmark, the automated extractor achieved accuracy 0.810, $F1 = 0.789$, and $\kappa = 0.616$, compared with reviewer–reviewer agreement of $\kappa = 0.506$. Feature-level extractor performance and independent second-annotator reliability are summarized in Figure 4.

3.3 Distribution of domain-level translational risk

Domain scores varied in both central tendency and spread across the cohort (Supplementary Figure S2; all on a 0–100 risk scale). Biological tractability had mean 40.8 (median 40, IQR 20–60); Clinical development mean 54.5 (median 56.6, IQR 32.4–75.5); Regulatory pathway mean 94.3 (median 100, IQR 100–100); Economic sustainability mean 50 (median 49.4, IQR 31.2–70.8); and Translation infrastructure mean 78.2 (median 81, IQR 70–94.2). Composite TRS had mean 63.6 (median 66.4, IQR

Table 1 Feature-level Type B extractor performance (n=332, ambiguous cases excluded), sorted by F1 ascending.

Feature	n	Precision	Recall	F1
consensus_guidance	19	0.500	0.714	0.588
multicenter_network	20	0.500	0.833	0.625
surrogate_endpoint_precedent	20	0.571	1.000	0.727
patient_registry	20	0.600	1.000	0.750
actionable_target	14	0.889	0.667	0.762
human_disease_model	20	0.700	1.000	0.824
orphan_designation_evidence	19	0.714	1.000	0.833
biomarker_in_trial	20	0.714	1.000	0.833
validated_outcome_measure	20	0.800	0.889	0.842
causal_molecular_basis	20	0.900	0.900	0.900
preclinical_rescue	20	1.000	0.833	0.909
biobank	20	1.000	0.833	0.909
natural_history_study	20	0.900	1.000	0.947
natural_history_consortium	20	0.900	1.000	0.947
patient_organization	20	1.000	1.000	1.000
in_vivo_model	20	1.000	1.000	1.000
expedited_designation_evidence	20	1.000	1.000	1.000

51.1–75.4, range 19.7–96.2). The Regulatory domain’s interquartile range was compressed at the ceiling (IQR 100–100, i.e. at least 75% of diseases scored at the maximum risk value), reflecting that reproducibly ascertainable regulatory-precedent evidence (orphan designation, expedited designation, surrogate-endpoint precedent) was confirmed for only a minority of diseases under the frozen retrieval protocol. This ceiling effect should be considered when interpreting the domain’s contribution to the bottleneck profiles reported below. Inter-domain correlations ranged from near-independent (Biological–Regulatory, $r=0.05$) to strongly correlated (Clinical–Economic, $r=0.85$), consistent with domains sharing some underlying documentation-driven variance while retaining distinguishable content.

3.4 Historical validation

Cross-validated discrimination for the primary historical outcome (Phase III advancement, n=100, 76 events) was AUC 0.822 (95% CI 0.719–0.913) for TRS alone; 0.771 (95% CI 0.654–0.881) for Evidence Coverage alone; 0.837 (95% CI 0.739–0.925) for TRS and Evidence Coverage jointly; and 0.838 (95% CI 0.742–0.923) for the five-domain multivariable model (Table 2, Figure 2). Corresponding Brier scores were 0.167, 0.193, 0.161, and 0.162, respectively; Evidence Coverage alone showed the weakest probabilistic performance. Univariate logistic regression confirmed the expected directions (Supplementary Figure S1): each 10-point increase in TRS was associated with an odds ratio of 0.334 (95% CI 0.203–0.550, $p=1.6 \times 10^{-5}$) for Phase III advancement — TRS is coded so that higher values represent greater modeled risk, so a value below 1 is the expected direction, not an inversion. Each 10-point increase in Evidence Coverage was associated with an odds ratio of 2.012 (95% CI 1.418–2.855,

$p=9.1 \times 10^{-5}$) for the same outcome. TRS and Evidence Coverage are highly correlated. Pearson $r(\text{TRS}, \text{Evidence Coverage}) = -0.941$ (Spearman $r = -0.938$), and both variables carried variance inflation factors of 8.72 in a joint model, indicating substantial collinearity. Better-documented rare diseases therefore tended to receive lower TRS values. A bootstrap comparison (1,000 replicates) favored TRS over Evidence Coverage alone (mean AUC difference 0.066, 95% CI 0.022–0.114, excluding zero), indicating that TRS is not simply a re-expression of documentation volume. However, adding Evidence Coverage to TRS produced no statistically significant incremental discrimination over TRS alone (mean AUC difference 0.016, 95% CI -0.010 to 0.078, crossing zero). Taken together, these results indicate that TRS carries discriminative signal beyond raw documentation quantity, but given $r=-0.94$, the two constructs are not cleanly separable, and joint-model coefficients should be interpreted with this collinearity in mind rather than as evidence of two independent risk axes.

Table 2 Cross-validated predictive performance by model specification.

Model	n	Events	AUC	95% CI	Brier score
TRS	100	76	0.822	0.719–0.913	0.167
Evidence Coverage	100	76	0.771	0.654–0.881	0.193
TRS + Evidence Coverage	100	76	0.837	0.739–0.925	0.161
Five Domains	100	76	0.838	0.742–0.923	0.162

Note: The Brier score is the mean squared error between predicted probabilities and observed binary outcomes; lower values indicate better probabilistic accuracy.

Domain-level comparisons were directionally consistent with the composite score: mean Biological, Clinical, Regulatory, Economic, and Infrastructure risk scores were all lower among diseases with Phase III advancement than among diseases without advancement (Figure 3).

3.5 Equal-weight versus fitted models

Comparing model specifications by bootstrap AUC difference (1,000 replicates each), the five-domain multivariable model showed a small improvement over TRS alone (mean difference 0.038, 95% CI 0.003–0.103), with the interval narrowly excluding zero. Given the width of this interval relative to the point estimate, the result provides only weak evidence of incremental value for the full five-domain specification over the equal-weight composite. **Domain-level coefficients (ridge-penalized).** An unpenalized multivariable logistic fit of the five domains showed quasi-complete separation (coefficient standard errors on the order of 10^5 to 10^7 for the Regulatory domain and intercept), making raw coefficient magnitudes and signs uninterpretable at this sample size ($n=100$, 24 non-events, 5 predictors). An L2-penalized (ridge) fit on standardized domain scores, with 1,000-replicate bootstrap 95% CIs (Table 3), found that only the Economic (coefficient -1.215, 95% CI -1.960 to -0.445) and Regulatory (coefficient -0.710, 95% CI -1.019 to -0.264) domains had bootstrap intervals excluding zero; Biological, Clinical, and Infrastructure domain coefficients were not distinguishable from

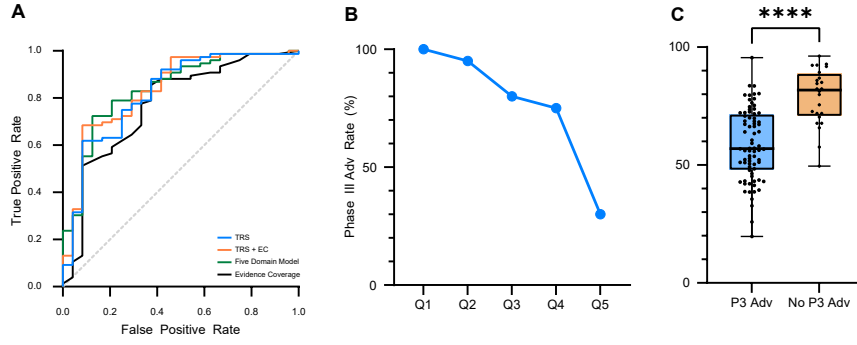


Fig. 2 Historical association between Translation Risk Score and subsequent Phase III advancement. (A) Cross-validated receiver operating characteristic curves for TRS, Evidence Coverage, TRS plus Evidence Coverage, and the five-domain model. Area under the curve values were 0.822 (95% CI 0.719–0.913), 0.771 (0.654–0.881), 0.837 (0.739–0.925), and 0.838 (0.742–0.923), respectively. (B) Phase III advancement across TRS quintiles. The 100-disease cohort was ranked from lowest to highest TRS and divided into five groups of 20 diseases; advancement occurred in 20/20 (100%), 19/20 (95%), 16/20 (80%), 15/20 (75%), and 6/20 (30%) diseases from Q1 through Q5, respectively. (C) TRS distributions according to observed Phase III advancement status. Median TRS was 56.95 (IQR 48.25–71.03) among diseases with advancement and 81.80 (IQR 71.53–87.50) among diseases without advancement. The two groups were compared using a two-sided Mann–Whitney test ($p < 0.0001$; ****). Phase III advancement denotes initiation of at least one Phase III trial during follow-up and does not imply trial success, efficacy, completion, or approval.

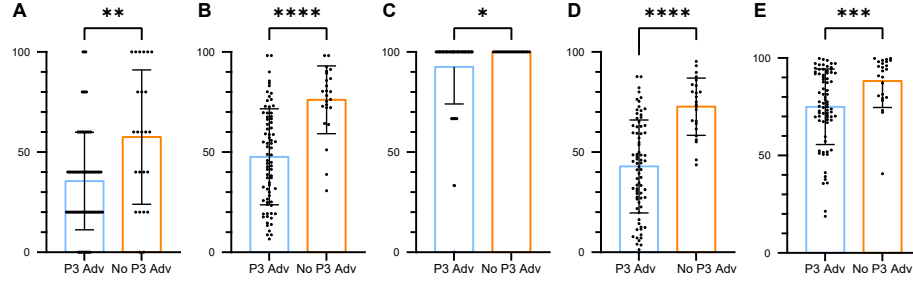


Fig. 3 Translational-domain risk scores by subsequent Phase III advancement status. Distributions of (A) Biological, (B) Clinical, (C) Regulatory, (D) Economic, and (E) Infrastructure domain scores are shown for diseases with Phase III advancement ($n=76$) and without Phase III advancement ($n=24$). Higher values indicate greater modeled translational risk. Mean domain scores were lower among diseases with Phase III advancement for Biological (35.5 vs 57.5), Clinical (47.6 vs 76.1), Regulatory (92.5 vs 100.0), Economic (42.8 vs 72.7), and Infrastructure (75.0 vs 88.2). Between-group comparisons were performed using two-sided Mann–Whitney tests. Significance symbols denote * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, and **** $p < 0.0001$. The Regulatory domain showed a pronounced ceiling effect, limiting between-group separation. Phase III advancement denotes initiation of at least one Phase III trial during follow-up.

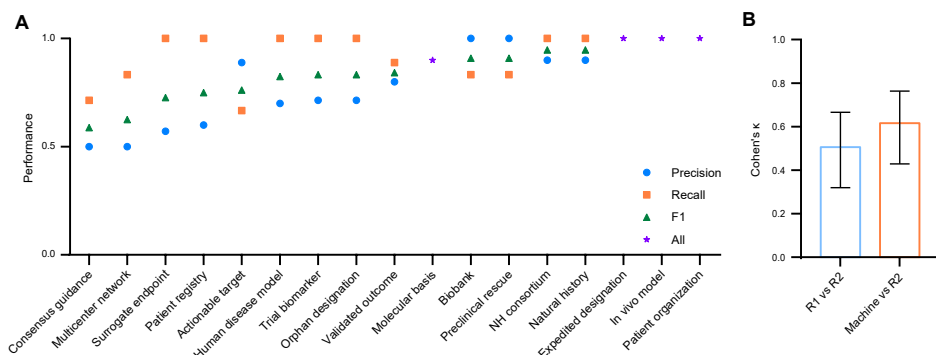


Fig. 4 Validation of documentary evidence extraction. (A) Feature-level precision, recall, and F1 score for the 17 prespecified Type B documentary evidence features. Of 340 reviewed decisions, 332 were evaluable after exclusion of eight reviewer-designated ambiguous cases; pooled performance was accuracy 0.873, precision 0.784, recall 0.901, F1 = 0.838, and Cohen’s κ = 0.735. **(B)** Independent second-annotator reliability in the blinded validation subset. Cohen’s κ was 0.506 (95% CI 0.320–0.666; n=83 non-ambiguous paired ratings) for Reviewer 1 versus Reviewer 2 and 0.616 (95% CI 0.429–0.763; n=84 non-ambiguous comparisons) for the machine extractor versus Reviewer 2. Raw reviewer–reviewer agreement was 74.1%; machine versus Reviewer 2 accuracy was 0.810, with F1 = 0.789.

zero at this sample size. This does not establish that Biological, Clinical, or Infrastructure risk is irrelevant to translational outcome — only that, jointly with the other four domains, their independent marginal contribution is not resolvable with n=100.

Table 3 Ridge-penalized standardized domain coefficients.

Domain	Standardized coefficient	95% bootstrap CI
Biological	−0.398	−0.978 to 0.135
Clinical	−0.421	−1.405 to 0.338
Regulatory	−0.710	−1.019 to −0.264
Economic	−1.215	−1.960 to −0.445
Infrastructure	0.119	−0.779 to 0.725

A leave-one-domain-out sensitivity analysis (recomputing the equal-weight composite from the remaining four domains) showed discrimination was not driven by any single domain: raw AUC ranged from 0.820 to 0.861 across the five leave-one-out composites, compared to 0.850 for the full five-domain equal-weight composite — no single domain’s exclusion produced a material discrimination loss.

3.6 Bottleneck profiles across diseases

The dominant modeled bottleneck for each disease (the domain with the highest risk score) was Regulatory for 79/100 diseases, Infrastructure for 12/100, and Biological for 9/100; no disease had Clinical or Economic as its single highest-risk domain (Table 4).

Caution on interpreting dominant-domain frequency: the Regulatory domain’s compressed, ceiling-heavy distribution (IQR pinned at the 100 maximum for most diseases) means it is mechanically likely to be selected as “dominant” whenever it reaches ceiling while other domains have not, independent of whether Regulatory risk is truly the most actionable constraint for that disease. This frequency table should be read as a description of which domain most often reaches its own ceiling in this cohort, not as a validated ranking of clinical importance, and is reported with that caveat rather than omitted.

Table 4 Dominant modeled bottleneck frequency by domain.

Domain	n diseases	% of cohort
Regulatory	79	79.0%
Infrastructure	12	12.0%
Biological	9	9.0%

3.7 Counterfactual Translation Risk

The frozen counterfactual analysis evaluated 370 eligible modifiable-feature scenarios across 25 diseases in the top-risk quartile, using the prespecified procedure described in Methods. Scenarios were ranked by modeled change in predicted probability of Phase III advancement. The largest modeled improvements clustered around expedited/orphan regulatory designation evidence and surrogate-endpoint precedent, consistent with the Regulatory domain’s ceiling-heavy distribution and bottleneck frequency: diseases with unconfirmed regulatory-precedent evidence have the most room for the fitted model to register a scenario improvement when that evidence is added (Figure 5). For the visceral myopathy case study, the same evidence definitions and scoring architecture were applied as in the cohort analysis, and the eligible single-feature counterfactual scenarios are presented in Supplementary Table S1.

Consistent with the prespecified counterfactual definition, these results are reported as model-based counterfactual scenario analyses for research prioritization, not as causal estimates of the effect of implementing the corresponding intervention (e.g., “under the fitted framework, adding surrogate-endpoint precedent evidence was associated with a change in predicted translational-advancement probability of the stated magnitude for the affected diseases,” not “pursuing a surrogate endpoint would increase the probability of Phase III success by that amount”).

4 Discussion

POROS was designed to characterize rare-disease translational readiness across several evidence domains rather than reduce readiness to a single marker of scientific maturity. In the frozen cohort of 100 diseases, higher TRS was associated with a lower probability of subsequent Phase III advancement. TRS alone produced a cross-validated AUC

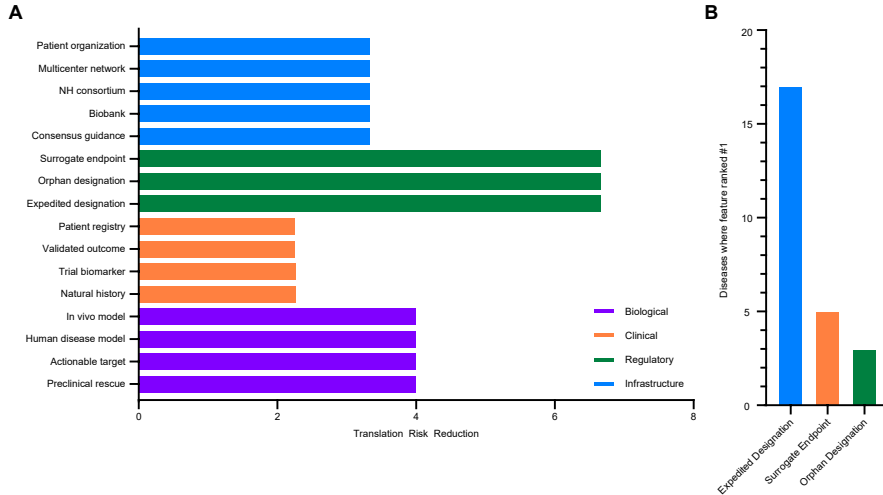


Fig. 5 Model-based counterfactual feature impact among diseases in the highest-risk quartile. (A) Mean modeled reduction in Translation Risk Score when a single eligible favorable evidence feature was added while all other inputs were held fixed. The largest modeled reductions were approximately 6.66 TRS points for expedited regulatory designation evidence, orphan designation evidence, and surrogate/intermediate endpoint precedent. **(B)** Frequency with which a feature ranked as the highest-leverage counterfactual among the 25 highest-risk diseases. Expedited regulatory designation evidence ranked first for 17 diseases, surrogate/intermediate endpoint precedent for five, and orphan designation evidence for three. These are fitted-model, single-feature counterfactual scenarios intended for research prioritization and should not be interpreted as causal effects of implementing the corresponding actions.

of 0.822, while advancement fell from 100% in the lowest-risk quintile to 30% in the highest-risk quintile. The five-domain model performed similarly, and adding Evidence Coverage to TRS produced only a small increase in discrimination. These findings are consistent with the broader rare-disease literature, where trial readiness depends on natural-history characterization, outcome measures, regulatory pathways, recruitment networks, and development infrastructure in addition to biological understanding [1–3].

The close relationship between TRS and Evidence Coverage is one of the most important limits on interpretation. Evidence Coverage alone showed substantial discrimination, and its correlation with TRS was strongly inverse ($r = -0.941$). The combined TRS plus Evidence Coverage model did not materially outperform TRS alone. Diseases with richer, more readily retrievable evidence also tended to have lower modeled translational risk, which is plausible because mature development fields often accumulate registries, validated outcome measures, biomarkers, regulatory precedents, and coordinated research networks. At the same time, this creates an identifiability problem. The present design cannot fully separate disease-intrinsic readiness from a field’s capacity to generate, index, and preserve documentary evidence. TRS should therefore be interpreted as a multidomain evidence-derived risk construct, not as a

pure measure of biological tractability. Reporting Evidence Coverage separately makes this dependence visible.

The domain-level results reinforce that caution. Biological, Clinical, Economic, and Infrastructure risk scores were lower on average among diseases that subsequently advanced to Phase III. The Regulatory domain showed a pronounced ceiling effect and was the dominant modeled bottleneck for 79 of 100 diseases. That frequency does not establish that regulatory barriers are the causal obstacle in 79% of rare diseases. Several regulatory features are sparse by construction and may only become observable after a disease field has already reached later stages of therapeutic development. Regulatory guidance likewise emphasizes coordinated progress in natural-history characterization, endpoint development, trial design, and regulatory evidence [7, 8]. Within POROS, the Regulatory result is better read as a recurrent evidence gap in the current scoring architecture than as an estimate of the causal effect of regulatory action.

Human validation places another boundary around the disease-level scores. The extractor performed well overall across 332 evaluable decisions, but feature-level precision and recall varied, so documentary-classification uncertainty propagates into downstream scores. In the blinded 85-item dual-annotator subset, Reviewer 1 and Reviewer 2 agreed on 74.1% of items, with Cohen’s $\kappa = 0.506$ (95% CI 0.320 to 0.666). Against Reviewer 2 as an independent reference, the machine achieved accuracy 0.810, F1 = 0.789, and $\kappa = 0.616$ (95% CI 0.429 to 0.763). Both human reviewers agreed while the machine differed in only 4 of 85 items; in 20 of 85 items, the human reviewers disagreed and the machine matched one of them. These patterns suggest that some disagreement arises from ambiguity in documentary boundaries rather than machine-specific error alone. They do not establish equivalence between automated and human review. Provenance, reviewable evidence snippets, and explicit handling of ambiguous cases remain important parts of the framework.

CTR analysis is intended for hypothesis generation. Among diseases in the highest-risk quartile, regulatory evidence features produced the largest modeled reductions in TRS. Expedited regulatory designation evidence ranked as the highest-leverage eligible feature for 17 of 25 diseases, surrogate or intermediate endpoint precedent for five, and orphan designation evidence for three. These scenarios show which missing evidence classes most strongly influence the frozen model. They do not estimate intervention effects. Adding a designation, endpoint precedent, registry, or disease model *in silico* does not show that creating that resource would cause a therapeutic program to advance. The scenarios hold other encoded inputs fixed and omit time, cost, feasibility, feature interactions, and the possibility that readiness features are themselves consequences of earlier development progress. CTR outputs should therefore be used as research-prioritization signals rather than causal recommendations [19].

Visceral myopathy illustrates the intended disease-level use of POROS. ACTG2-associated visceral myopathy has an established molecular basis [20, 21], yet its modeled profile remains constrained by gaps outside causal-gene identification. The case study shows how strong mechanistic evidence can coexist with limited longitudinal cohorts, validated outcome measures, trial infrastructure, regulatory precedent, or other resources needed for therapeutic development. POROS does not predict the

fate of a specific visceral-myopathy program. It identifies where modeled evidence gaps occur and preserves the provenance needed to inspect them.

Several features of the study design improve interpretability. Historical evidence was frozen before outcome ascertainment, reducing direct temporal leakage. Documentary classifications retain provenance rather than only final binary labels, missing evidence is tracked separately from modeled risk, and extractor performance was assessed against both the primary human review and an independent second annotator. The descriptive TRS also retains its equal-weight structure rather than being retrospectively reweighted to maximize association with the observed outcome. These choices improve auditability, although they cannot remove limitations inherent to the public evidence ecosystem.

The study also has substantial limitations. The cohort contains 100 diseases and only 24 non-events, which restricts the stability and complexity of multivariable models. Ridge penalization and bootstrap confidence intervals reduce instability but cannot remove the limitations imposed by sample size [17, 22, 23]. Temporal separation reduces direct information leakage but does not eliminate confounding, survivorship bias, historical changes in disease nomenclature or registry practices, or reverse directionality between readiness features and therapeutic investment. Phase III advancement was defined as initiation of at least one Phase III trial, so the endpoint does not measure trial completion, efficacy, regulatory approval, commercial success, or patient benefit. Cohort-relative normalization also means that TRS expresses position within the frozen reference cohort rather than an absolute biological scale. Finally, the strong dependence between TRS and Evidence Coverage limits claims that the score captures a construct fully independent of documentation intensity.

All reported discrimination and probabilistic-performance estimates reflect internal validation. Cross-validated performance does not establish transportability to new diseases, later time periods, or different evidence ecosystems [24–26]. External temporal and disease-level validation is therefore the next major test of POROS. Future studies should examine whether the TRS gradient persists in independently assembled cohorts, evaluate calibration as evidence coverage changes, and test whether individual domains retain incremental value after accounting for documentary availability. Larger cohorts with more non-events would support more stable multivariable estimation and stronger assessment of domain-specific effects. Alternative weighting schemes and assumptions about missing evidence should also be examined. Until those tests are completed, POROS is best used as an auditable research-prioritization framework that organizes heterogeneous evidence and generates testable hypotheses about translational bottlenecks.

5 Conclusions

POROS provides an evidence-derived framework for decomposing rare-disease translational readiness into biological, clinical, regulatory, economic, and infrastructure domains. In the frozen 100-disease cohort, lower TRS was associated with greater subsequent Phase III advancement, with an internally cross-validated AUC of 0.822. Human review supported the performance of the documentary evidence extractor

while also showing meaningful reviewer variability. Counterfactual analyses identified evidence gaps associated with the largest modeled reductions in TRS, but these scenarios remain non-causal. Interpretation is constrained by the modest cohort size, strong TRS–Evidence Coverage collinearity, a ceiling-heavy Regulatory domain, reliance on public documentary evidence, and use of Phase III initiation rather than trial success as the outcome. External cohorts with more non-events are needed to assess transportability, evaluate calibration, and test the stability of domain-specific effects. At this stage, POROS is best treated as an auditable research-prioritization framework rather than a clinical or regulatory decision instrument.

Supplementary information. The following supporting information accompanies this manuscript: Supplementary Figure S1, univariate associations with Phase III advancement; Supplementary Figure S2, POROS score-component and evidence-completeness analyses; and Supplementary Table S1, visceral-myopathy counterfactual scenarios.

Acknowledgements. The author thanks Lydia Lu for serving as the independent secondary reviewer for the blinded documentary evidence validation. Generative AI tools, including ChatGPT (OpenAI), were used during manuscript preparation for language refinement, code troubleshooting, and editorial review. All scientific content, analyses, interpretations, and final wording were reviewed and verified by the author, who takes full responsibility for the manuscript.

Declarations

Funding. This research received no external funding.

Competing interests. The author declares no competing interests.

Ethics approval and consent to participate. Not applicable. This study analyzed publicly available, non-identifiable disease-level, clinical-trial, and documentary data and did not involve human participants or animals.

Consent for publication. Not applicable.

Data availability. The frozen disease-level analysis dataset, feature dictionary, and derived statistical and counterfactual-analysis outputs supporting this study are publicly available in Zenodo at <https://doi.org/10.5281/zenodo.22168501>. Original third-party source documents are not redistributed; the public repository contains derived study variables and analysis outputs.

Materials availability. Not applicable.

Code availability. Source code for the POROS/RDTI analysis pipeline is available at <https://github.com/thegrapeslayer/poros-platform>; the manuscript-matched software release is archived at <https://doi.org/10.5281/zenodo.22168470>.

Author contribution. H.S. was responsible for conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing–original draft preparation, writing–review and editing, visualization, and project administration. The author read and approved the final manuscript.

Abbreviations.	AUC	Area under the receiver operating characteristic curve
	CI	Confidence interval
	CTR	Counterfactual Translation Risk
	F1	Harmonic mean of precision and recall
	P3	Phase III
	POROS	Provenance-linked translational-risk framework
	ROC	Receiver operating characteristic
	TRS	Translation Risk Score
	VIF	Variance inflation factor

References

- [1] Kempf L, Goldsmith JC, Temple R. Challenges of developing and conducting clinical trials in rare disorders. *Am J Med Genet A*. 2018;176(4):773–783. <https://doi.org/10.1002/ajmg.a.38413>.
- [2] Austin CP, Cutillo CM, Lau LPL, Jonker AH, Rath A, Julkowska D, et al. Future of Rare Diseases Research 2017–2027: an IRDiRC perspective. *Clin Transl Sci*. 2018;11(1):21–27. <https://doi.org/10.1111/cts.12500>.
- [3] Lumsden JM, Urv TK. The Rare Diseases Clinical Research Network: a model for clinical trial readiness. *Ther Adv Rare Dis*. 2023;4:26330040231219272. <https://doi.org/10.1177/26330040231219272>.
- [4] Miller KL, Fermaglich LJ, Maynard J. Using four decades of FDA orphan drug designations to describe trends in rare disease drug development: substantial growth seen in development of drugs for rare oncologic, neurologic, and pediatric-onset diseases. *Orphanet J Rare Dis*. 2021;16(1):265. <https://doi.org/10.1186/s13023-021-01901-6>.
- [5] Fermaglich LJ, Miller KL. A comprehensive study of the rare diseases and conditions targeted by orphan drug designations and approvals over the forty years of the Orphan Drug Act. *Orphanet J Rare Dis*. 2023;18(1):163. <https://doi.org/10.1186/s13023-023-02790-7>.
- [6] Sinha MS, Stern AD, Rai AK. Four decades of orphan drugs and priorities for the future. *N Engl J Med*. 2024;391(2):100–102. <https://doi.org/10.1056/NEJMp2402816>.
- [7] US Food and Drug Administration.: Rare diseases: natural history studies for drug development. Draft guidance for industry. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/rare-diseases-natural-history-studies-drug-development>. Accessed 29 Aug 2026. Silver Spring (MD): US Department of Health and Human Services.
- [8] US Food and Drug Administration.: Rare diseases: considerations for the development of drugs and biological products. Guidance for industry.

<https://www.fda.gov/regulatory-information/search-fda-guidance-documents/rare-diseases-considerations-development-drugs-and-biological-products>. Accessed 29 Aug 2026. Silver Spring (MD): US Department of Health and Human Services.

- [9] Kitahara K, Kano S. Assessing rare disease understanding: a novel disease readiness level framework. *Orphanet Journal of Rare Diseases*. 2025;20:628. <https://doi.org/10.1186/s13023-025-04135-y>.
- [10] Zarin DA, Tse T, Williams RJ, Califf RM, Ide NC. The ClinicalTrials.gov results database — update and key issues. *N Engl J Med*. 2011;364(9):852–860. <https://doi.org/10.1056/NEJMsa1012065>.
- [11] Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20(1):37–46. <https://doi.org/10.1177/001316446002000104>.
- [12] Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. New York: Chapman and Hall; 1993.
- [13] Harrell FE Jr. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. 2nd ed. Cham: Springer; 2015.
- [14] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29–36. <https://doi.org/10.1148/radiology.143.1.7063747>.
- [15] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128–138. <https://doi.org/10.1097/EDE.0b013e3181c30fb2>.
- [16] Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev*. 1950;78(1):1–3. [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2).
- [17] Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*. 1970;12(1):55–67. <https://doi.org/10.1080/00401706.1970.10488634>.
- [18] Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Stat Med*. 2002;21(16):2409–2419. <https://doi.org/10.1002/sim.1047>.
- [19] Hernán MA. The C-word: scientific euphemisms do not improve causal inference from observational data. *Am J Public Health*. 2018;108(5):616–619. <https://doi.org/10.2105/AJPH.2018.304337>.

- [20] Lehtonen HJ, Sipponen T, Tojkander S, Karikoski R, Jarvinen H, Laing NG, et al. Segregation of a missense variant in enteric smooth muscle actin γ -2 with autosomal dominant familial visceral myopathy. *Gastroenterology*. 2012;143(6):1482–1491.e3. <https://doi.org/10.1053/j.gastro.2012.08.045>.
- [21] Wangler MF, Gonzaga-Jauregui C, Gambin T, Penney S, Moss T, Chopra A, et al. Heterozygous de novo and inherited mutations in the smooth muscle actin (ACTG2) gene underlie megacystis-microcolon-intestinal hypoperistalsis syndrome. *PLoS Genet*. 2014;10(3):e1004258. <https://doi.org/10.1371/journal.pgen.1004258>.
- [22] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49(12):1373–1379. [https://doi.org/10.1016/S0895-4356\(96\)00236-3](https://doi.org/10.1016/S0895-4356(96)00236-3).
- [23] Riley RD, Snell KIE, Ensor J, Burke DL, Harrell J Frank E, Moons KGM, et al. Minimum sample size for developing a multivariable prediction model: PART II — binary and time-to-event outcomes. *Stat Med*. 2019;38(7):1276–1296. <https://doi.org/10.1002/sim.7992>.
- [24] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230. <https://doi.org/10.1186/s12916-019-1466-7>.
- [25] Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. *BMC Med*. 2023;21(1):70. <https://doi.org/10.1186/s12916-023-02779-w>.
- [26] Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. <https://doi.org/10.1136/bmj-2023-078378>.